Network Working Group	F.J. Baker
Internet-Draft	Cisco Systems
Intended status: Informational	July 02, 2011
Expires: January 03, 2012	

Routing a Traffic Class
draft-baker-fun-routing-class-00

Abstract

This note addresses the concept of routing a traffic class. This has many possible implementations, IGP and BGP, and link state as well as distance vector. The fundamental impetus is the question raised in RFC 3704 and shim6 of exit routing, the question raised by Mike O'Dell of source/destination routing, and the "fish" problem, raised in many networks, in which distinct traffic classes that could conceivably use the same route predictably use different routes. Instead of handling these as "destination routing with a twist", the paper looks at the matter systemically.

Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.
Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet- Drafts is at http://datatracker.ietf.org/drafts/current/.
Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."
This Internet-Draft will expire on January 03, 2012.

Copyright Notice

Table of Contents

1. Introduction

This note addresses the concept of routing a traffic class. This has many possible implementations, IGP and BGP, and link state as well as distance vector. The fundamental impetus is the question raised in [RFC3704] and shim6 of exit routing, the question raised by Mike O'Dell of source/destination routing, and the "fish" problem, raised in many networks, in which distinct traffic classes that could conceivably use the same route predictably use different routes. Instead of handling

these as "destination routing with a twist", the paper looks at the matter systemically.

1.1. Scope

The question of implementation of IPv4 or IPv6 routes is moot; the algorithms discussed here can be used for either. Examples, however, will be drawn from IPv6 *[RFC2460]* and will presume its addressing architecture *[RFC4291]*.

1.2. Structure of the paper

The paper looks first at the fundamental concept in Section 2. It then goes on to imagine an IS-IS-like *[RFC5308]* [RFC6119] implementation. Unfortunately, this really can't be implemented in IS-IS by adding a TLV, which is our usual approach, due to the changed concept of a metric. However, given the changed concepts of a metric and a TLV, it shows how the same exchange and calculation algorithms can be used to build such a routing protocol. It also looks at a RIP-like *[RFC2080]* algorithm that includes a sequence number (derived from AODV *[RFC3561]*) to simplify count-to-infinity-related problems. It stops short of a BGP implementation, although one modeled on the distance vector model would make sense.

2. The fundamental concept of routing a class

This section introduces the fundamental concepts involved in routing a traffic class. These include the definitions of

* a "traffic class", which is a set-theoretic definition - all traffic matching a constraint,

* a "metric", which is an administrative number applied to a class of traffic crossing an interface, and

* a "route announcement", which is the accumulation of information regarding the routing of a traffic class as modified by various metrics en route.

In addition, it describes the fundamental algorithm applied in modifying a route announced by a predecessor node using a metric for announcement on the interface implied.

2.1. Define: "traffic class"

A "class" of traffic, in any routing protocol, is the selector that is used to identify a traffic stream for the purpose of routing. For traditional internet routing protocols like RIP, OSPF, IS-IS, EIGRP, or

BGP, a "traffic class" is "the traffic destined to a stated prefix". In the context of this design, a traffic class is the traffic that goes

*to a destination prefix, which may be a default route (::/0), a host route (any /128 prefix), or anything in between,

*from a source prefix, which may be a default route (::/0), a host route (any /128 prefix), or anything in between, and

*using one of a set of DSCP *[RFC2474]* values.

The set of DSCP values includes and "any DSCP". "Any DSCP" has the obvious meaning: we are not really looking at the DSCP in traffic classification.
A protocol could also identify a route as a "null route"; this is like any other route, but specifically directs that traffic matching the traffic class is to be dropped.
A traffic class is represented in this document as

*{destination, source, {list of DSCPs}|any|none}

Examples of common traffic classes include:

Standard Default Route: {::/0, ::/0, any}

Default Route using a source prefix {::/0, source, any}

Default route used only for a QoS Traffic Class: {::/0, ::/0, {list of DSCPs}}

Examples of QoS traffic classes are found in the Configuration Guidelines for DiffServ Service Classes *[RFC4594]* and related documents [RFC5127] and [RFC5865].

2.2. Define: "metric"

A "metric" is an attribute of an interface, and associates a traffic class with a administrative value used in comparison of routes. In this document, it is represented as

*{{destination, source, {DSCP}|any|none}, administrative value}

and should be understood as "the metric for traffic from source to destination using DSCP".
For example, if an interface is available to all VoIP traffic, one of the metrics on an interface might be

*{{::/0, ::/0, {EF}}, administrative value}

2.3. Define: "route announcement"

A route announcement is identical to a metric in structure, but is semantically different. Where a metric is an attribute of an interface, a route is an entry in the route table calculated by the routing protocol, and is used in making forwarding decisions.
The calculation of a route follows the following logic:

1. Inputs to the route calculation are a route announcement (which may be derived from an interface metric in the case of local routes, or received from a neighboring route for remote routes), and a metric associated with an interface.

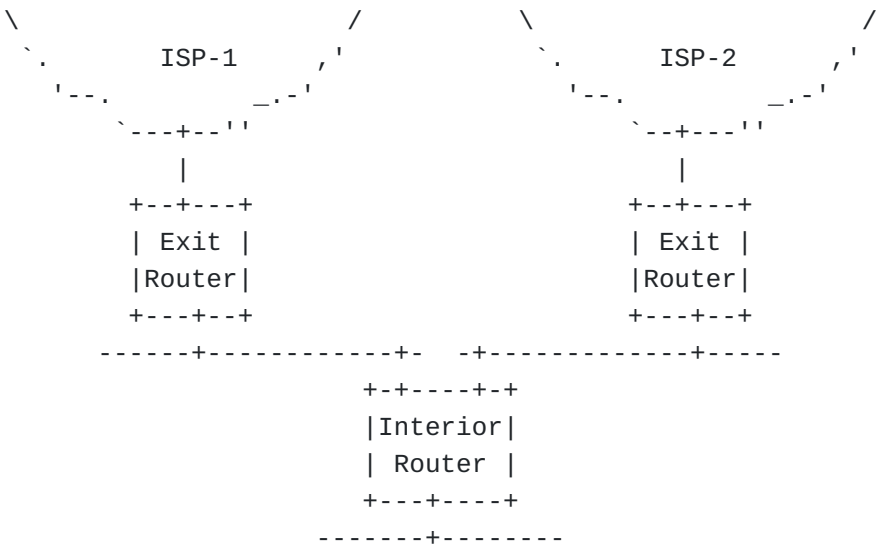
2. The source, destination, and set of DSCPs of the route announcement and the metric are intersected to generate the traffic class for the resulting route.

3. The resulting route's administrative value is the sum of the original route's administrative value and the interface's administrative value.

For example, in Figure 1, if the prefix allocated by ISP-1 to a network is 2001:db8:1::/48 and the prefix allocated by ISP-2 to the network is 2001:db8:2::/48, the exit routers for the network might advertise into the network the default routes

*{{2000::/3, 2001:db8:1::/48, any}, 1} ("unicast traffic using source addresses in 2001:db8:1::/48 for any application"), and

*{{::/0, 2001:db8:2::/48, any}, 2} ("all traffic using source addresses in 2001:db8:2::/48 for any application").



One edge case is the case in which a route intersected with the available metrics yields the null set. In such a case, the resulting route cannot be used as no traffic matches it.
If an interior distance vector router in the network receives

*{{2000::/3, 2001:db8:1::/48, any}, 1} and

*{{::/0, 2001:db8:2::/48, any}, 2}

from its upstream router, and on a given interface has a metric

*{{::/0, ::/0, {EF}}, 5} ("any VoIP traffic"),

it would validly infer the routes

*{{2000::/3, 2001:db8:1::/48, {EF}}, 6} ("unicast VoIP traffic using source addresses in 2001:db8:1::/48") and

*{{::/0, 2001:db8:2::/48, {EF}}, 7} ("all VoIP traffic using source addresses in 2001:db8:2::/48").

It would install the routes it received into its own route table, and advertise the calculated routes to neighboring routers.
Note that there is no question of a unicast RPF or other forms of Ingress Filtering *[RFC2827]* required in such a network. If a host spoofs a source address in another routing domain and sends a datagram upstream, there is no route in the network "from" that routing domain, and the router has no idea what to do with it. In such cases the router would silently drop the traffic (it might be nice to respond with an ICMP, but to whom?); it should of course maintain appropriate counters and/or logs. Similarly, since traffic is routed according to both its source and destination addresses, traffic using a source address in 2001:db8:2::/48 would have no chance of existing to ISP-1. It is still possible for traffic from outside to come in, however, as such routes would have a source prefix of ::/0 or 2000::/3.

2.4. Route Precedence

Precedence in selection of routes is based on intersection, and follows the rule "most specific first". This is a generalization of the "longest match first" rule used in destination routing. That may require, in generating the Forwarding Information Base (FIB) from the Routing Information Base (RIB), that we calculate the least intersection of two route announcements.
In calculating routes, either one route's traffic class is a subset of another's, or they mutually incompatible. If one is a subset of the other, we consider the superset to be "less specific" and the subset to be "more specific"; the most specific matching traffic class rules. If the most specific option is in fact multiple traffic classes none of which are subsets of the others, we calculate the intersections of

those routes for the purpose of comparison, and apply the rule to the resulting route announcements.
For example, consider the two routes

*{{2000::/3, ::/0, any}, 1} and

*{{::/0, 2000::/3, any}, 5}.

They are not comparable because neither is a subset of the other. We therefore calculate the intersection, which is a choice between

*{{2000::/3, 2000::/3, any}, 1} and

*{{2000::/3, 2000::/3, any}, 5}.

Given that choice, the metric clearly selects {{2000::/3, 2000::/3, any}, 1}. So we install in the RIB the three routes

*{{2000::/3, 2000::/3, any}, 1},

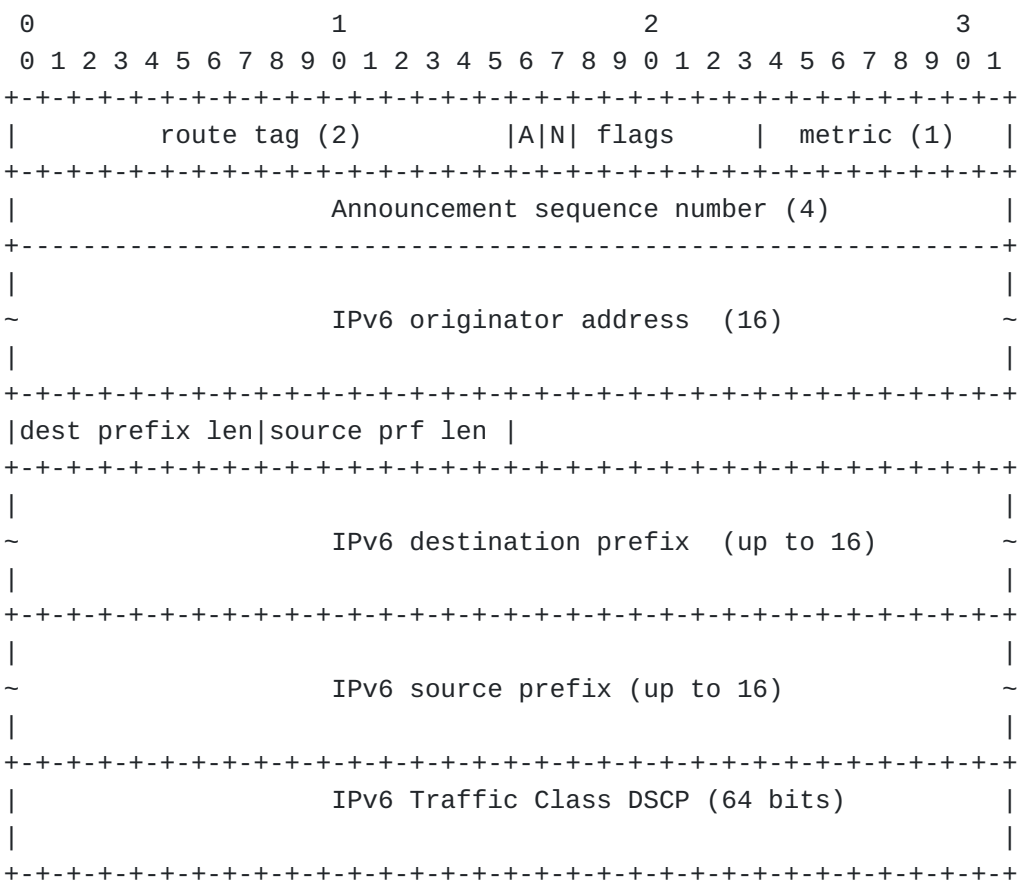
*{{2000::/3, ::/0, any}, 1}, and

*{{::/0, 2000::/3, any}, 5}.

The first is used by unicast traffic, as it is the most specific that matches traffic from and to addresses in 2000::/3. The second is used, for example, with traffic that is from addresses whose most significant three bits are not 001 and to an address in 2000::/3. The third is used for traffic that is to addresses whose most significant three bits are not 001 but are from addresses in 2000::/3.

3. Sketch of a distance vector implementation

A distance vector implementation would operate much as described in Section 2. In addition, however, we might take advantage of an algorithm used in AODV *[RFC3561]*) to simplify count-to-infinity-related problems.
To this end, we use the mechanisms and algorithms of [RFC2080] with certain modifications.
A Route Table Entry (RTE), in this model, might be structured as in Figure 2.



The fields are:

Route Tag: The Route Tag is as defined in [RFC2080].

A: If 1, any DSCP applies, and the IPv6 Traffic Class DSCP mask is absent. If 0, the IPv6 Traffic Class DSCP is present.

N: if 1, a null route; traffic in this traffic class and not in more explicit match SHOULD be dropped.

flags: unspecified in this version of the specification. Set to 0 on transmission and ignored on receipt. Future versions may specify additional flags.

Metric: The Metric is as defined in [RFC2080].

Announcement sequence number: 0..FFFFFFFF, follows the rules for a sequence number specified in [RFC3561].

IPv6 originator address: One of the global addresses of the router originating this RTE, presumably the one used on the relevant interface, which is in control of the sequence number. See [RFC3561] for considerations and algorithms.

Destination Prefix Length: 0..128

Source Prefix Length:
0..128

IPv6 destination prefix: The significant bits of the prefix in question. Occupies ceiling(destination prefix length/8) bytes.

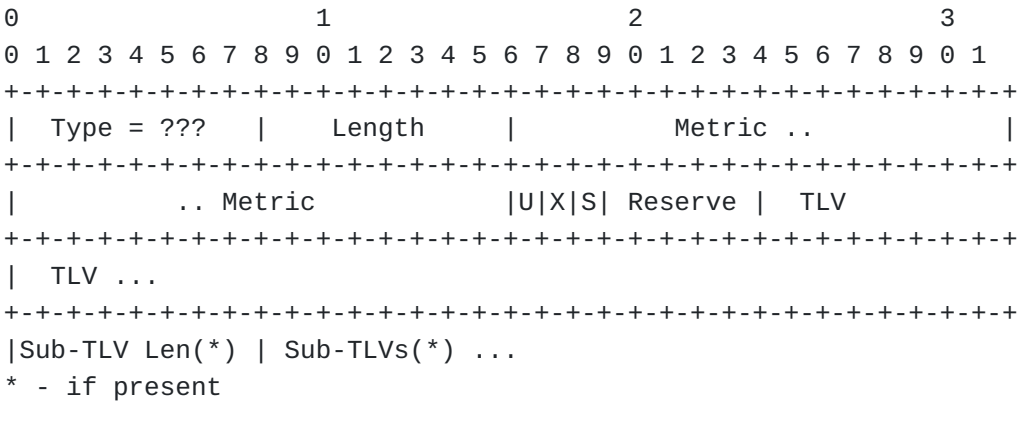
IPv6 source prefix: The significant bits of the prefix in question. Occupies ceiling(source prefix length/8) bytes.

IPv6 Traffic Class DSCP: if A=0, not present; if A=0, the most significant bit is numbered 0, and indicates that the traffic class includes traffic with a DSCP of 000000; if A=1, not present.

Distribution and calculation of routes follows [RFC2080], including split horizon (an announcement received from a router on the interface should not be reannounced to the same router). and poison reverse (which should be used on triggered updates but not regular announcements).
Elimination of stale routing information of routes follows the algorithm with the sequence number specified in [RFC3561].
Intersection of route announcements with interface metric data is as described in Section 2.

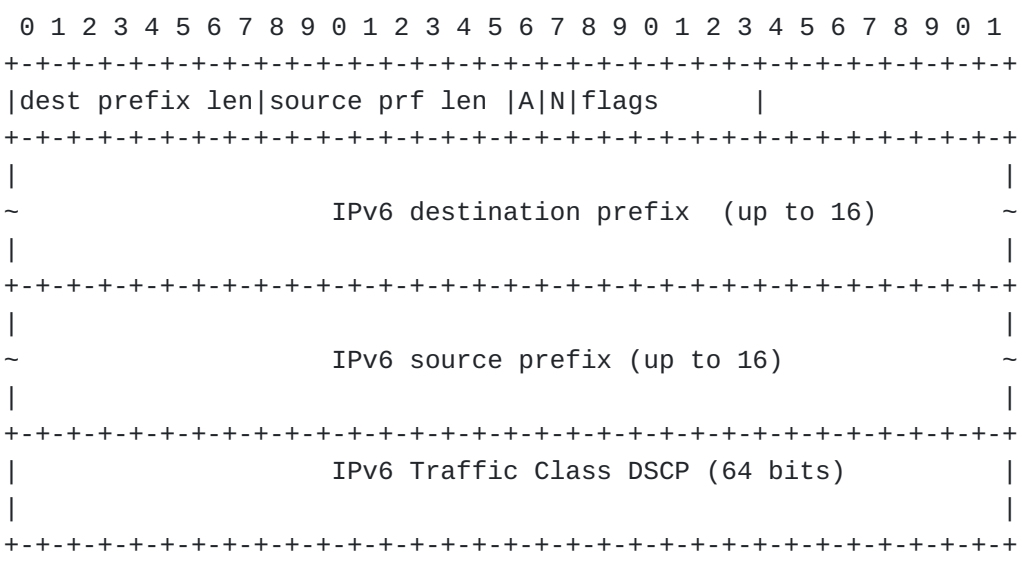
4. Sketch of a link state implementation

A link state (SPF) implementation would also operate much as described in Section 2, although the algorithm operates in memory as opposed to being distributed.
To this end, we use the mechanisms and algorithms of [RFC5308] with certain modifications.
IS-IS structures its network as a lattice of routers that lead one to sets of hosts identified by "TLVs". A TLV, in this model, might be structured as in Figure 3.



* - if present

TLV or sub-TLV format:



The fields are:

U: up/down bit

X: external original bit

S: subtlv present bit

A: If 1, any DSCP applies, and the IPv6 Traffic Class DSCP mask is absent. If 0, the IPv6 Traffic Class DSCP is present.

N: if 1, a null route; traffic in this traffic class and not in more explicit match SHOULD be dropped.

flags: unspecified in this version of the specification. Set to 0 on transmission and ignored on receipt. Future versions may specify additional flags.

Destination Prefix Length:
0..128

Source Prefix Length: 0..128

IPv6 destination prefix: The significant bits of the prefix in question. Occupies ceiling(destination prefix length/8) bytes.

IPv6 source prefix: The significant bits of the prefix in question. Occupies ceiling(source prefix length/8) bytes.

IPv6 Traffic Class DSCP: if A=0, not present; if A=0, the most significant bit is numbered 0, and indicates that the traffic class includes traffic with a DSCP of 000000; if A=1, not present.

Distribution and calculation of routes follows [RFC5308]. Intersection of route announcements with interface metric data is as described in Section 2.

5. IANA Considerations

At this point, this memo asks the IANA for no new parameters and gives the IANA no instructions. As development progresses, that might change.

6. Security Considerations

Security issues in routing protocols such as these are the same as in other distance vector and link state routing protocols, and need to be mitigated in the same ways. Since this paper looks primarily at the algorithms for route calculation, those issues are largely ignored. If a protocol such as described in Section 3 or Section 4 is implemented, however, care must be taken to ensure the integrity of communications between routers and their mutual authentication.

7. Acknowledgements

Dana Blair reviewed the initial version of the draft.

8. Change Log

Initial Version: Sat Mar 26 12:25:46 CET 2011

9. References

9.1. Normative References

[RFC2119]	Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
[RFC2460]	

	Deering, S.E. and R.M. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.
[RFC4291]	Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

9.2. Informative References

[RFC2080]	Malkin, G. and R. Minnear, "RIPng for IPv6", RFC 2080, January 1997.
[RFC2474]	Nichols, K., Blake, S., Baker, F. and D.L. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.
[RFC2827]	Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, May 2000.
[RFC3561]	Perkins, C., Belding-Royer, E. and S. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing", RFC 3561, July 2003.
[RFC3704]	Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, March 2004.
[RFC4594]	Babiarz, J., Chan, K. and F. Baker, "Configuration Guidelines for DiffServ Service Classes", RFC 4594, August 2006.
[RFC5127]	Chan, K., Babiarz, J. and F. Baker, "Aggregation of DiffServ Service Classes", RFC 5127, February 2008.
[RFC5308]	Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, October 2008.
[RFC5865]	Baker, F., Polk, J. and M. Dolly, "A Differentiated Services Code Point (DSCP) for Capacity-Admitted Traffic", RFC 5865, May 2010.
[RFC6119]	Harrison, J., Berger, J. and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, February 2011.

Author's Address

Fred Baker Baker Cisco Systems Santa Barbara, California 93117 USA
EMail: fred@cisco.com